Lead Scoring Case Study – Summary

Steps Followed

- Importing necessary libraries
- Importing the provided dataset
- Data Wrangling
- Exploratory Data Analysis (Variables Inspection)
- Data Preparation
- Model Building (Logistic Regression)
- Model Evaluation (Logistic Regression Metrics)
- Model Testing
- Model Inference
- Conclusion based on our results

Data Wrangling:

- Import dataset
- Go through the entire dataset and make key observations.
- Check overall dimensions of the dataset.
- Check column formats and correct any irregularities found in dataset.
- Check for any NULL values present in the dataset.
- Deal with NULL values by imputing those rows or replacing with mean or median values.

Exploratory Data Analysis (EDA):

- Data imbalance was checked, and the ratio was found to be 1:1.6 (converted to not converted).
- Univariate and multivariate categorical analysis was made on all features, and count plots were displayed.
- Columns with high data imbalance were dropped.
- Univariate and multivariate numerical analysis was carried out on all numerical columns, and a pair plot and heatmap were plotted.
- Boxplot analysis was made to handle and treat outliers present.

Data Preparation:

- **Binary level categorical columns were already mapped to 1 / 0 in previous steps**
- **Created dummy features (one-hot encoded) for categorical variables – Lead Origin, Lead Source , Do not email, Last Activity, Specialization, Current occupation, Tags ,City , A free copy of Mastering the Interview, Last Notable activity.**
- **Splitting Train & Test Sets -** 70:30 % ratio was chosen form the split
- **Feature scaling** - Standardization method was used to scale the features
- **Checking the correlations -** Predictor variables which were highly correlated with each other were dropped.

Model Building:

- The dataset has many dimensions and features.
- This can reduce model performance and increase computation time.
- **Recursive Feature Elimination** (RFE) is important to select only important columns.
- Manual Feature Reduction was used by dropping variables with p-value greater than 0.05.
- Model 3 looks stable after 3 iterations with significant p-values within the threshold (p-values < 0.05)
- There is no sign of **multicollinearity** with VIFs less than 5.
- Model 3 will be the final model used for Model Evaluation and predictions.

Model Evaluation:

- The final trained model had an accuracy score of 91%, Precision score of 89%, F1 score of 93%, and ROC curve area of 97% after choosing the optimal cut off at 0.35 from the graph of accuracy, sensitivity, and specificity.
- lead score was assigned for the trained data.

Metrics	Scores
Accuracy score	0.915
F1- score	0.931
Precision score	0.891
Recall score	0.915

Model Testing:

- The built model was then tested on the test data where we got an accuracy score of 82%, sensitivity of 80%, and an F1 Score of 77%. Hence the model was stable.
- Lead score was then assigned to the tested data.

Conclusion:

- Landing Page Submissions and Lead Add Form lead to more conversions.
- Conversions are higher for leads from Google, Organic Search, Direct Traffic, and Referrals.
- SMS and Email marketing leads have higher conversions.
- Finance, HR, Marketing, Operations, and Banking sector leads tend to convert more.
- "Better Career Prospects" option for career outcome leads to higher conversions.
- Leads spending more time on the website tend to convert more.
- Reducing website bounce rate can increase customer engagement time and conversions.
- Lead Add Form generates qualifying leads and should be used across key areas.
- Sales team should focus on working professionals for higher conversions.
- Leads with a Lead Score >0.35 tend to convert more and model accuracy score is 91%.